NATIONAL CANCER INSTITUTE

Division of Cancer Control
& Population Sciences

Statistical Weights and Methods for Analyzing HINTS Data

HINTS Data Users Conference
January 21, 2005

William W. Davis, Ph.D.
Richard P. Moser, Ph.D.
National Cancer Institute

HINTS Survey Carried Out by Westat

- List of telephone exchanges purchased
- Exchanges and numbers sampled using random digit dialing (RDD)
  - Screens out unwanted exchanges (e.g., business exchanges)
  - Exchanges with high minority representation were over-sampled (HINTS stratification)
- For more information see L. Rizzo’s document on our website
  - “NCI HINTS Sample Design and Weighting Plan”

HINTS Statistical Weight

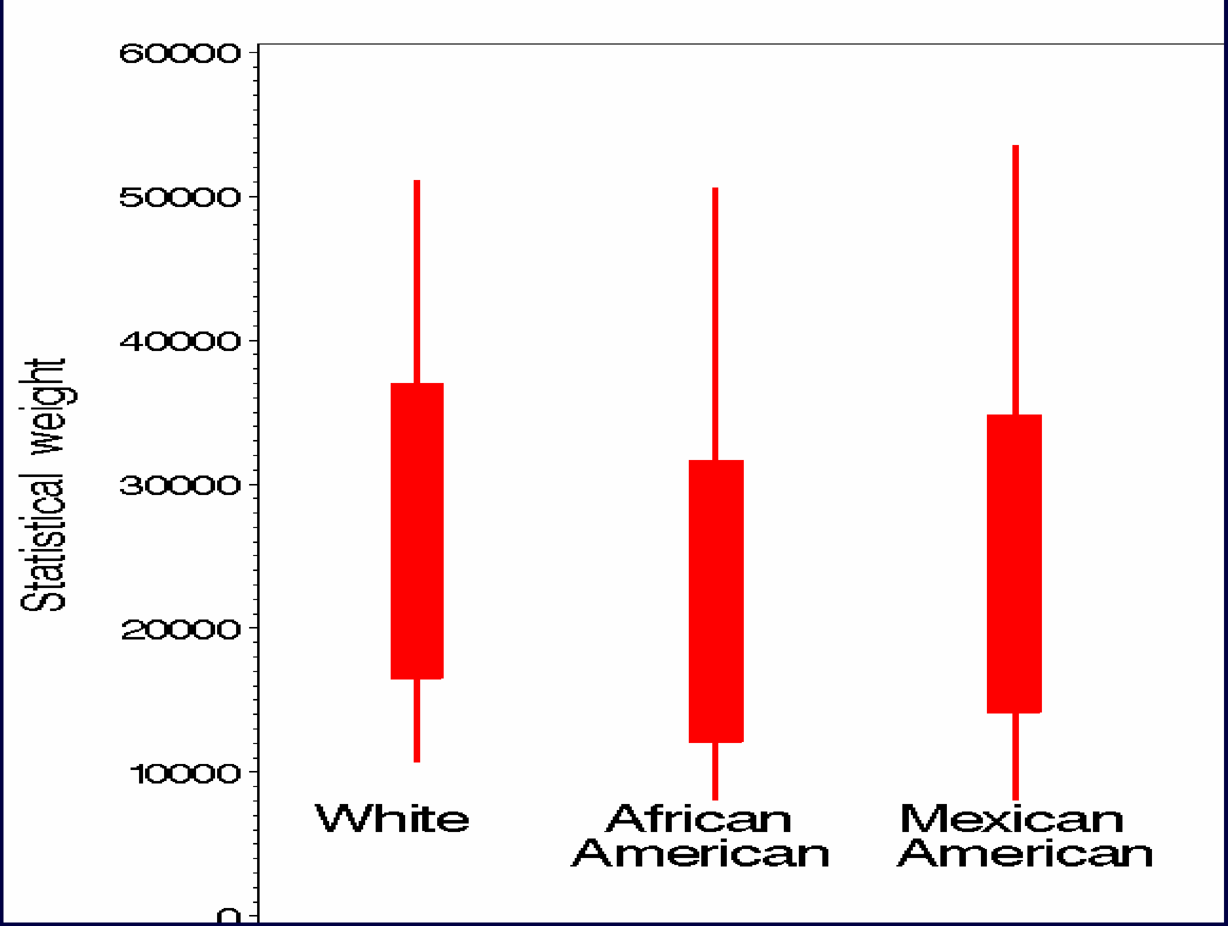
- Statistical weight:
    - Sampled person represents this many in the population
- HINTS Statistical weights derived from
    - Selection probabilities,
    - Number of telephones in the household
    - Response rates
    - Post-stratification adjustment

HINTS: Race Ethnicity

Race Eth	N	%		Wgt N	Wgt %	Diff %
Hispanic	764	12.0%		23,340,239	11.1%	**0.9%**
White	4276	67.1%		143,031,482	68.3%	-1.1%
Afr Amer	716	11.2%		20,905,523	10.0%	**1.3%**
Others	312	4.9%		12,028,337	5.7%	-0.8%
Missing	301	4.7%		10,148,812	4.8%	-0.1%
Total	**6369**	**100.0%**		**209,454,391**	**100.0%**	

Reflects the planned oversampling of minority exchanges.

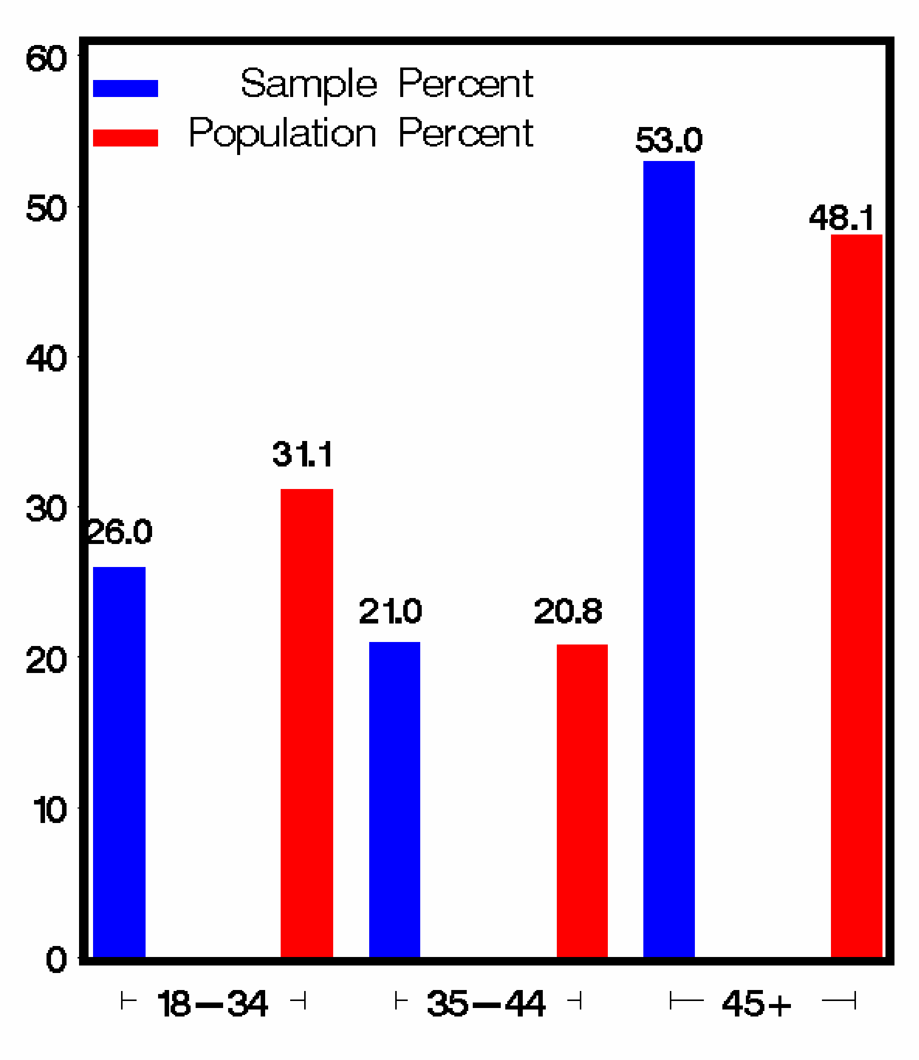
Minorities Were Oversampled: Boxplot of Statistical Weights

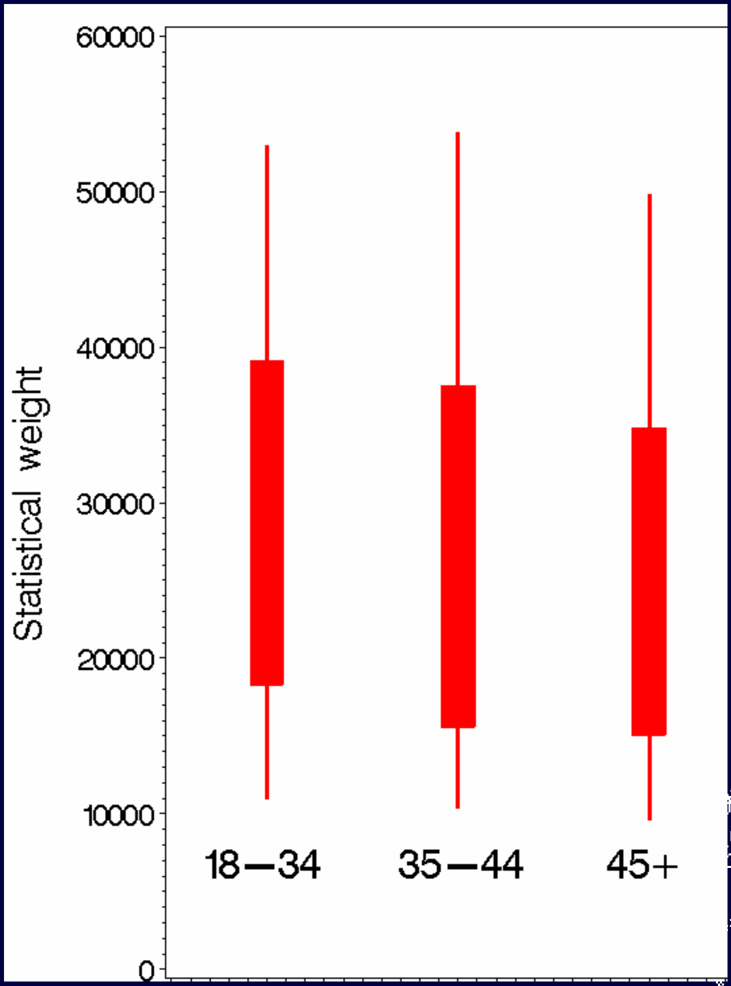


Older Folks Participated at a Higher Rate

NATIONAL CANCER INSTITUTE

Division of Cancer Control & Population Sciences

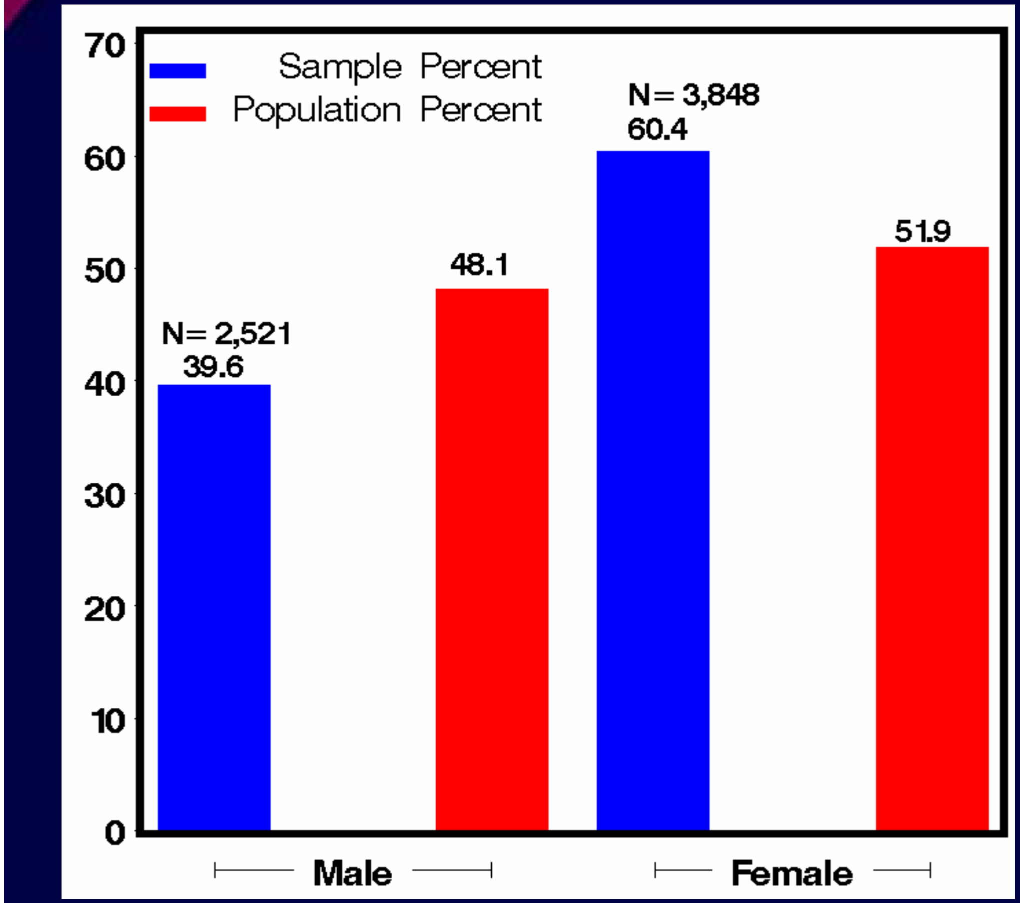


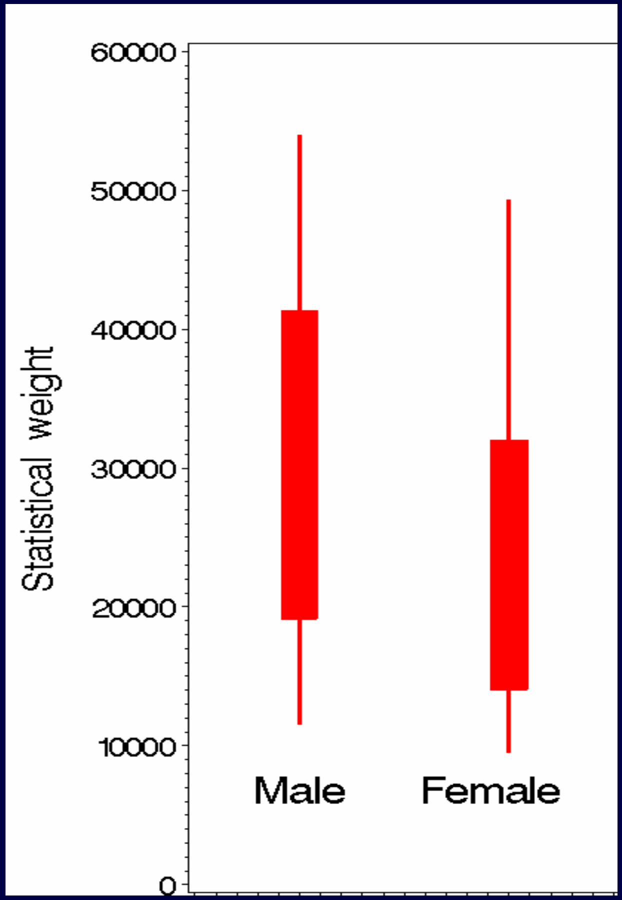


Females Participated at a Higher Rate

National Cancer Institute

Division of Cancer Control & Population Sciences

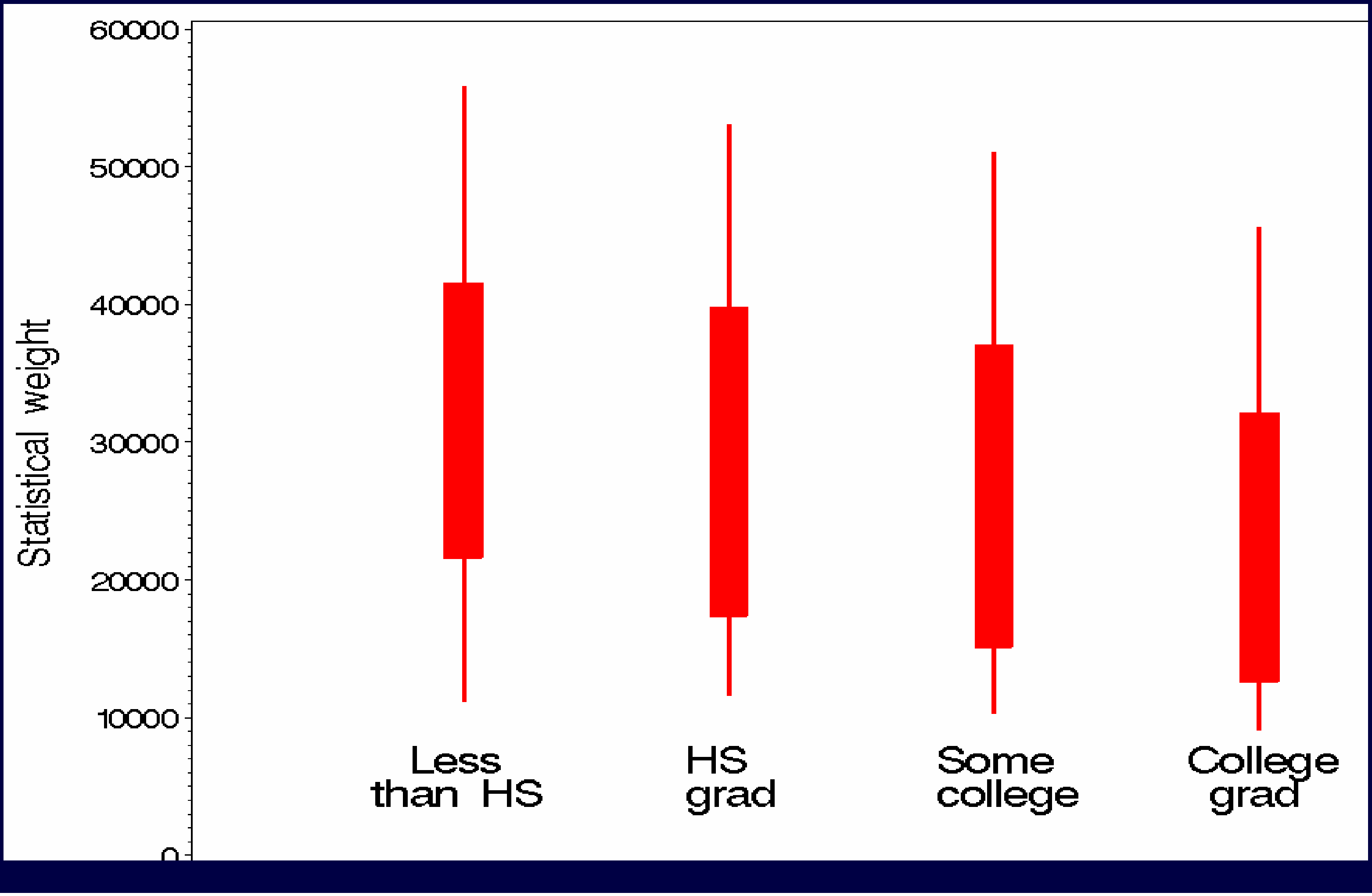




NATIONAL CANCER INSTITUTE

Division of Cancer Control & Population Sciences

Participation Increased with Education



HINTS: Weighted vs. Unweighted Analyses

Unweighted HINTS analyses would have

- Too many African Americans and Hispanics
- Too many 45+ and too few 18-34 year olds
- Too many females and too few males
- Too many people with high education

Variance/Bias Tradeoff for Mean

Estimate	Mean	Confidence Interval
Unweighted	$\overline{y}_u$	$\overline{y}_u \pm 1.96\sigma\left(\overline{y}_u\right)$
Weighted	$\overline{y}_w$	$\overline{y}_w \pm 1.96\sigma\left(\overline{y}_w\right)$

- The unweighted mean is biased
- The weighted mean has a larger variance

$$\sigma\left(\overline{y}_w\right) = \sigma\left(\overline{y}_u\right)\sqrt{1+CV^2}$$

HINTS Design Effect

$$\sigma\left(\overline{y}_w\right)=\sigma\left(\overline{y}_u\right)\sqrt{1+CV^2}$$

- CV is the coefficient of variation of the stat. weights
- *1+CV*^{2} is called the design effect
- CIs are 17-31% larger due to the weights
- Small price to pay for correct centering

Restriction	CV	$(1+CV^2)^{1/2}$
African American females	0.84	1.31
Hispanic males	0.61	1.17

Replicate weights

- What are replicate weights?
  - HINTS 50 replicate weights (fwgt1-fwgt50) were obtained by deleting 1/50th of the subjects in the full sample (and re-weighting)
- Why do we need replicate weights?
  - Used to estimate the variance of estimates obtained from the full sample -- for example a mean or a regression coefficient
- For more information see the SUDAAN manual or
  - Korn, E.L. and Graubard, B.I. (1999). Analysis of Health Surveys. John Wiley, p. 29.

Examples of HINTS Weights

Sub	fwgt	fwgt1	fwgt2
1	14,367	14,693	14,837
2	109,694	111,069	111,021
3	14,767	0	14,859
4	18,467	19,301	0

Full sample (fwgt) and 2 replicate weights for 4 sampled people
First two subjects are in both replicates while other two are not
The sum of each column of weights is the same – 209,454,391

Jackknife Estimate of Variance

Full sample estimate	$\hat{\theta}$
Replicate estimate	$\hat{\theta}_i$
Jackknife estimate of variance	$Var(\hat{\theta}) = \frac{49}{50}\sum_{i=1}^{50}\left(\hat{\theta}_i - \hat{\theta}\right)^2$

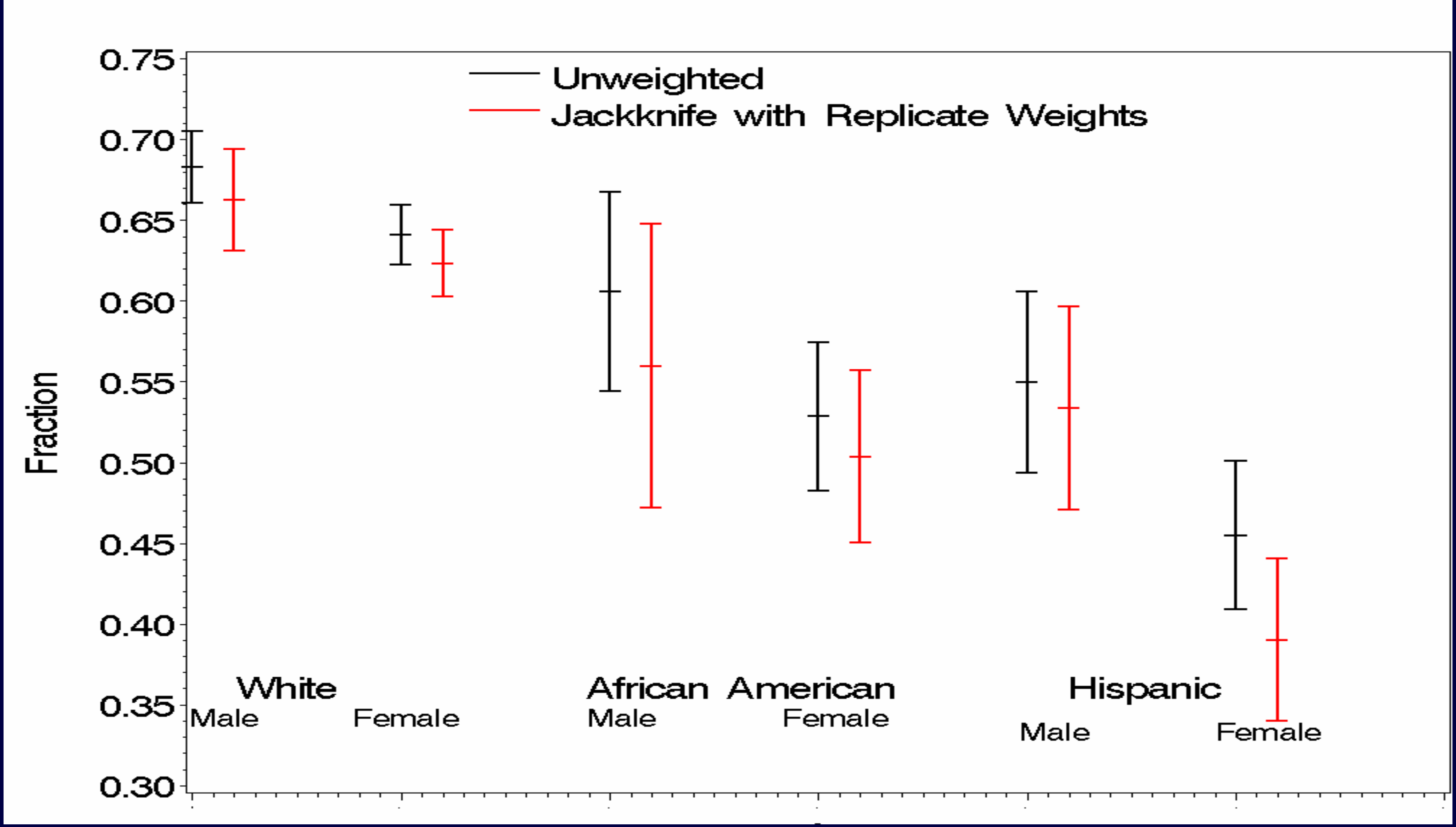
HINTS Example

- I'm going to read you a list of organizations. Before being contacted for this study, had you ever heard of:

  a. The National Institutes of Health?

- Estimate population proportion (and give 95% confidence intervals (CIs))
  - by race-ethnicity/gender

Heard of NIH? – 95% CIs



Division of Cancer Control & Population Sciences

SAS and SUDAAN Procedures

Analysis type	SAS	SUDAAN
	Not designed for survey analysis	Designed for survey analysis
Mean	MEANS	DESCRIPT
Crosstab	FREQ	CROSSTAB
Multiple regression	REG or GLM	REGRESS
Logistic regression	LOGISTIC	RLOGIST

Comparing Results with Logistic Regression: SAS vs. SUDAAN

- SAS Proc Logistic (unweighted; weighted)
- SUDAAN Proc Rlogist (weighted)
  - Proc Logistic (Standalone)
- Model: Internet= Age Education Race
  - Where:
    - Outcome= Ever accessed internet (Yes=1)
    - Age (continuous)
    - Education (4 levels; ref= LT High School)
    - Race (5 levels; ref= Hispanic)

SUDAAN Syntax

```
proc rlogist data=test design=jackknife;
weight fwgt;
jackwgts fwgt1-fwgt50/adjjack=.98;
class educ newrace;
reflev educ=1 newrace=1;
model internet= spage educ newrace ;
run;
```

Note: Design, Weight, Jackwgts, and Adjjack statements are used regardless of procedure in SUDAAN

Results: Education

Value Some College vs. LT High School (ref)	SAS Proc Logistic Unweighted	SAS Proc Logistic Weighted	SUDAAN Proc Rlogist
Log Odds (Odds)	2.23 (9.30)	2.04 (7.71)	2.04 (7.71)
Standard Error (β)	0.11	0.10	0.15
95% CI (β)	(2.01, 2.45)	(1.85, 2.23)	(1.74, 2.34)

Note: Larger standard error and corresponding CI with SUDAAN

Results: Race

Value White vs. Hispanic (ref)	SAS Proc Logistic Unweighted	SAS Proc Logistic Weighted	SUDAAN Proc Rlogist
Log Odds (Odds)	0.62 (1.86)	1.04 (2.84)	1.04 (2.84)
Standard Error	0.09	0.10	0.15
95% CI	(0.43, 0.81)	(0.85, 1.24)	(0.74, 1.35)

Summary

- HINTS unweighted estimates are biased
- HINTS weights vary by race/ethnicity, gender, age and education
- HINTS replicate weights can be used to obtain valid confidence intervals
- We compare weighted and unweighted analyses using means and logistic regression